Model Facts

Putting AI Ethics into Practice and Establishing Trust with End Users

@ the 2025 INFORMS Analytics+ Conference

Presentation by:

Jessica Zhu

Jessica Zhu, Michel Cukier, Joseph Richardson, Nutrition facts, drug facts, and model facts: putting AI ethics into practice in gun violence research, *Journal of the American Medical Informatics Association*, Volume 31, Issue 10, October 2024, Pages 2414–2421, https://doi.org/10.1093/jamia/ocae102

https://github.com/jhzsquared/model_facts

Agenda

- Motivation
- Background
- Our Approach: Model Facts
- Moving Forward

Motivation

[1]

[1] https://simonsinek.com/golden-circle/

A Public Health Crisis: Gun Violence in America

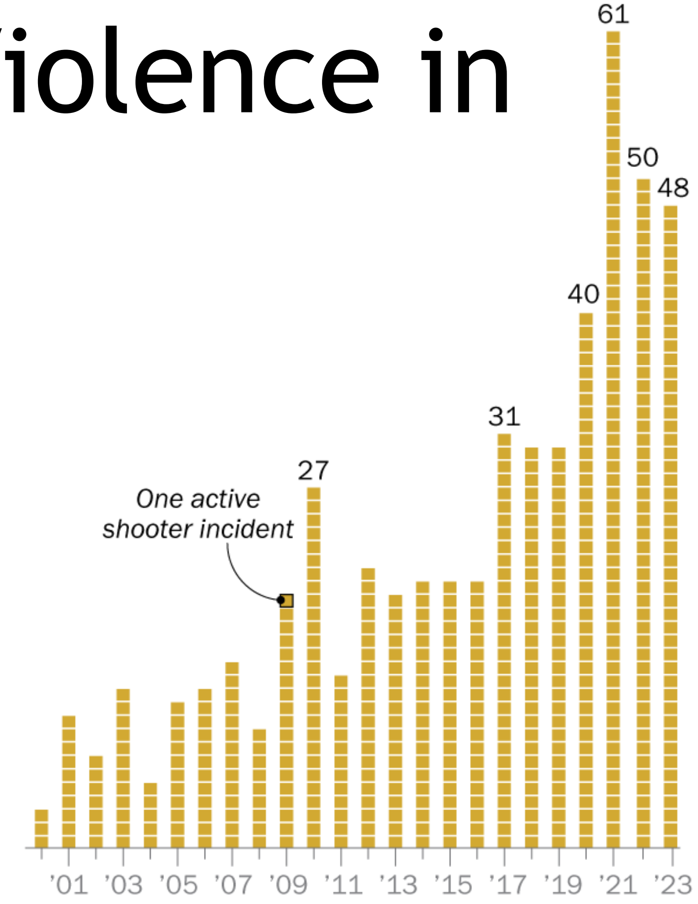
- Gun violence costs the US $557 billion annually [1]
- Firearm injuries are the leading cause of death for children, adolescents, and young Black men in the US
- Research funding was blocked by the Dickey Amendment until FY2020
- As of 2023, we found 11 machine learning publications with applications to gun violence research

How do we ethically apply ML to support Gun Violence researchers and avoid propagating social biases?

PROGRESS

Prevent Gun Violence: Research Empowerment Strategies & Solutions

Number of active shooter incidents, by year



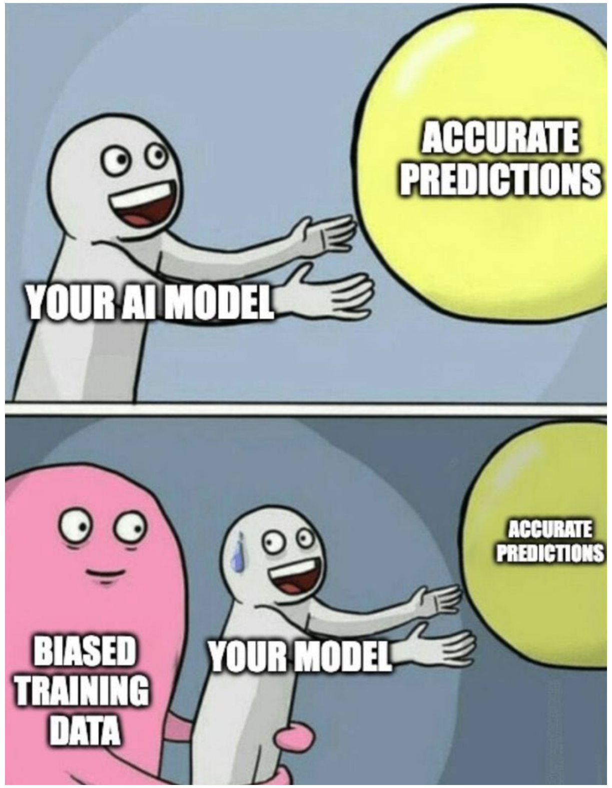
Note: "Active shooter incidents" are defined by the FBI as "one or more individuals actively engaged in killing or attempting to kill people in a populated area."
Source: Federal Bureau of Investigation. Data last accessed on Feb. 21, 2025.

PEW RESEARCH CENTER

[2]

[1] https://hms.harvard.edu/news/business-case-reducing-gun-violence
[2] https://www.pewresearch.org/short-reads/2025/03/05/what-the-data-says-about-gun-deaths-in-the-us/sr_25-03-05_gun-deaths_5/

Trust Starts with Transparency



[1]

[2]

[1] https://www.linkedin.com/posts/tommychavez_ethicaltech-algorithmicbias-activity-7111046195526979584-Nys2/
[2] https://www.instagram.com/data.science.beginners/p/DCoeFODTMAE/

Historical Misuse

1972 **Today in History: July 25, Tuskegee Syphilis Study exposed** [1]

AUTOPSIES

2016 [2]

Bernard Parker, left, was rated high risk

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

2019 [3]

ARTIFICIAL INTELLIGENCE

A US government study confirms most face recognition systems are racist

By Karen Hao

December 20, 2019

The New York Times [4]

2023

Black Artists Say A.I. Shows Bias, With Algorithms Erasing Their History

Tech companies acknowledge machine-learning algorithms can perpetuate discrimination and need improvement.

Forbes | Subscribe: Less than $1.50/wk [5]

2024

NYPD ShotSpotter Gunshot Detection Is Wildly Inaccurate, New Study Finds

By Lars Daniel, Contributor. Lars Daniel covers digital evidence and forensic... Follow Author

Dec 05, 2024, 01:07pm EST

LDN

2025

When algorithms decide who gets a loan: The fraught fight to purge bias from AI [6]

[1] https://apnews.com/today-in-history/july-25
[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[3] https://www.technologyreview.com/2019/12/20/79/ai-face-recognition-racist-us-government-nist-study/
[4] https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html
[5] https://www.forbes.com/sites/larsdaniel/2024/12/05/new-study-nypd-shotspotter-gunshot-detection-is-wildly-inaccurate/
[6] https://www.londondaily.news/when-algorithms-decide-who-gets-a-loan-the-fraught-fight-to-purge-bias-from-ai/

Background

Original Label

Nutrition Facts

Serving Size 2/3 cup (55g)
Servings Per Container 8

Amount Per Serving	
Calories 230	Calories from Fat 70
	% Daily Value*
Total Fat 8g	**12%**
Saturated Fat 1g	**5%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 160mg	**7%**
Total Carbohydrate 37g	**12%**
Dietary Fiber 4g	**16%**
Sugars 12g	
Protein 3g	
Vitamin A	10%
Vitamin C	8%
Calcium	20%
Iron	45%

* Percent Daily Values are based on a 2,000 calorie diet. Your Daily Value may be higher or lower depending on your calorie needs.

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

New Label

Nutrition Facts

8 servings per container
Serving size 2/3 cup (55g)

Amount per serving	
Calories	**230**
	% Daily Value*
Total Fat 8g	**10%**
Saturated Fat 1g	**5%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 160mg	**7%**
Total Carbohydrate 37g	**13%**
Dietary Fiber 4g	**14%**
Total Sugars 12g	
Includes 10g Added Sugars	**20%**
Protein 3g	
Vitamin D 2mcg	10%
Calcium 260mg	20%
Iron 8mg	45%
Potassium 240mg	6%

* The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

[1]

[1] https://www.fda.gov/media/97999/download

Dataset Nutrition Label

- Goal: "make it easier for practitioners to quickly assess the viability and fitness of datasets they intend to train AI algorithms on." [1]
- Draws inspiration from Datasheets for Datasets [2]

[1] https://datanutrition.org/labels/
[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92. https://doi.org/10.1145/3458723
[3] Holland, S. , Hosny, A. , Newman, S. , Joseph, J. , & Chmielinski, K. (2020). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. In D. Hallinan , R. Leenes , S. Gutwirth & P. De Hert (Ed.). Data Protection and Privacy: Data Protection and Democracy (pp. 1-26). Oxford: Hart Publishing. Retrieved March 14, 2025, from http://dx.doi.org/10.5040/9781509932771.ch-001

DATA NUTRITION PROJECT

Get more details

75% COMPLETENESS

What is this label?

The Dataset Nutrition Label enhances context, contents, and legibility of datasets. Information about this Dataset Nutrition Label, a standard of documentation that describes how to understand and use this dataset.

Consulted by	Name T. at DNP
First published on	Apr 6, 2021
Last updated on	Jan 6, 2022
Label version	Version 2.0

Hide

Public

Studies of Human Cognition with Neural Language Models

Preview data | Download PDF

Description

Using crowdsourcing framework MTurk, researchers first collect recalled stories and summaries from workers, then provide these summaries to other workers who write imagined stories. Finally, months later, researchers collect a retold version of the recalled stories from a subset of recalled authors.

Keywords

Language | Memory | Cognition | Computer science | Machine learning

How to use it?

Safe | Caution | Risky | Unknown

Intended Use
Examining cognitive processes of remembering and imagin... **Read more**

Restrictions on Use
Change this copy with restrictions on use ... **Read more**

Known Use
Recollection versus imagination: Exploring memory an... **Read more**

Do Not Use
Predicting characteristics of specific U.S. sub-populations... **Read more**

About the dataset

People

Created by	M. Sap, Y.Choi & 4 others
Owned by	M. Sap, Y.Choi & 4 others
Maintained by	M. Sap, Y.Choi & 4 others

Inference Risks

At-a-glance

About humans	Technical quality review	Upstream sources	Ethical review	Update frequency
Yes	Yes	Zero	Yes	Not Known

Technical information

Creation date	Jan 20, 2022
Format	Tabular, csv
Instances	6,854 narratives
Version	V3
Collection process	Self-reported

Useful links

Data dictionary	V3README.txt
License	CDLA Permissive 2.0
GitHub link	URL goes here

Data Values

What values are in each column?

- Manipulating data
- Other imputation
- Labeled data & protocols
- Missing or removed data
- (How) is raw data accessible

Number of issues: Risky 2, Safe 2, Unknown 1

Feature selection

Which columns were chosen and why?

- Includes confidential data
- Proxy characteristics
- Data definitions
- Domain-specific knowledge
- Consent given

Number of issues: Risky 2, Safe 3, Unknown 0

Representation

Which rows were included and why?

- Subpopulations identified
- Identities included
- Extrapolating trends
- Other issues

Number of issues: Risky 4, Safe 0, Unknown 0

Upstream Data

Are there known issues in datasets upstream?

- Dataset pulling from upstream sources
- Data processed
- Data collected
- Others

Number of issues: Risky 0, Safe 4, Unknown 0

General risks

Describe what this section includes

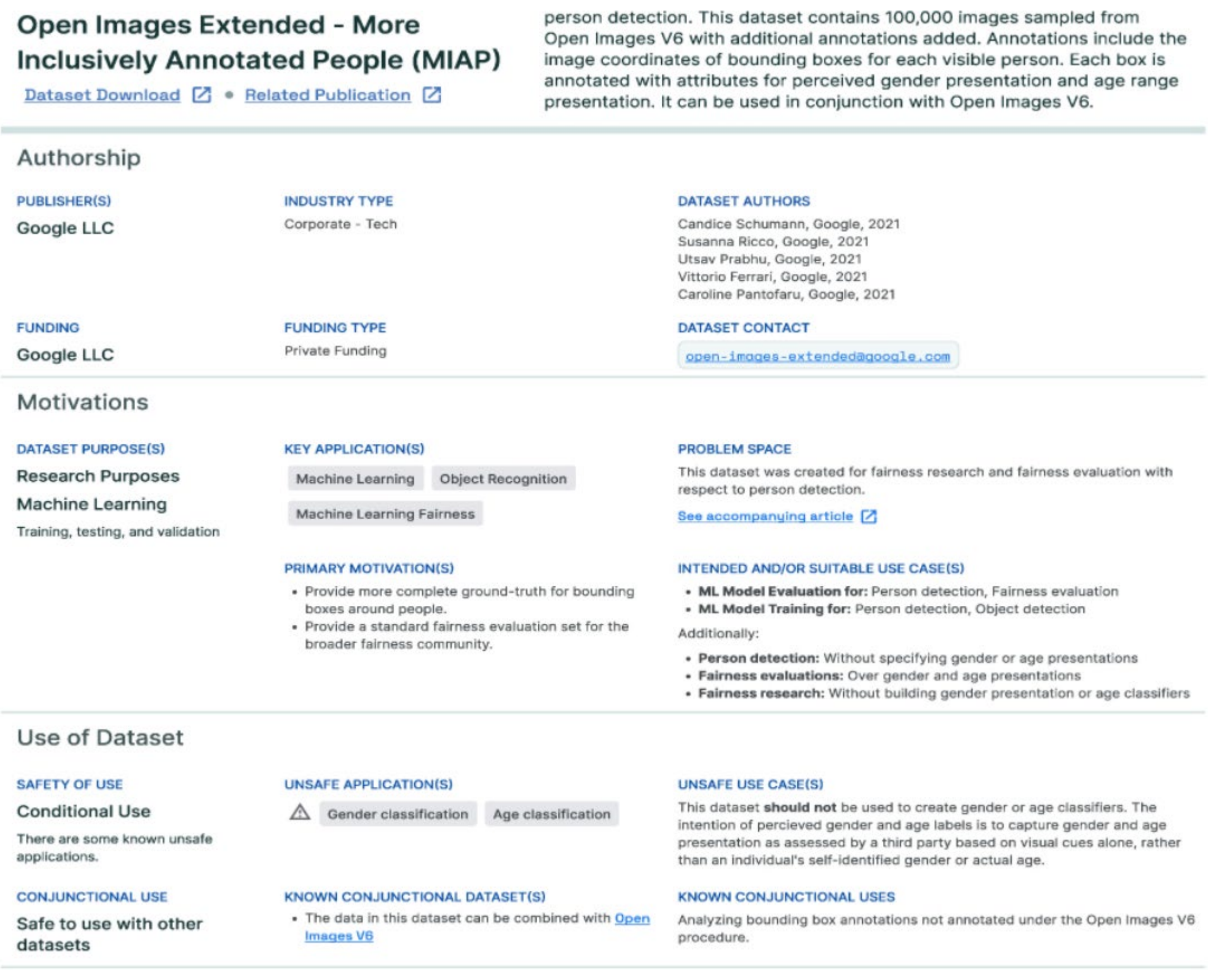
- Updates
- Funding
- Decisions on the individual lebel
- Sensitive content warning
- Ethical review

Number of issues: Risky 0, Safe 5, Unknown 0

[1] https://datanutrition.org/labels/

Data Cards Playbook

- Goal: Support continuous and contextual data transparency [2]
- Structured summaries of essential facts
- Goes through method, labels, attributes, access



Open Images Extended - More Inclusively Annotated People (MIAP)

Dataset Download • Related Publication

person detection. This dataset contains 100,000 images sampled from Open Images V6 with additional annotations added. Annotations include the image coordinates of bounding boxes for each visible person. Each box is annotated with attributes for perceived gender presentation and age range presentation. It can be used in conjunction with Open Images V6.

Authorship

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

DATASET AUTHORS
Candice Schumann, Google, 2021
Susanna Ricco, Google, 2021
Utsav Prabhu, Google, 2021
Vittorio Ferrari, Google, 2021
Caroline Pantofaru, Google, 2021

FUNDING
Google LLC

FUNDING TYPE
Private Funding

DATASET CONTACT
open-images-extended@google.com

Motivations

DATASET PURPOSE(S)
Research Purposes
Machine Learning
Training, testing, and validation

KEY APPLICATION(S)
Machine Learning · Object Recognition · Machine Learning Fairness

PROBLEM SPACE
This dataset was created for fairness research and fairness evaluation with respect to person detection.
See accompanying article

PRIMARY MOTIVATION(S)
- Provide more complete ground-truth for bounding boxes around people.
- Provide a standard fairness evaluation set for the broader fairness community.

INTENDED AND/OR SUITABLE USE CASE(S)
- **ML Model Evaluation for:** Person detection, Fairness evaluation
- **ML Model Training for:** Person detection, Object detection

Additionally:
- **Person detection:** Without specifying gender or age presentations
- **Fairness evaluations:** Over gender and age presentations
- **Fairness research:** Without building gender presentation or age classifiers

Use of Dataset

SAFETY OF USE
Conditional Use
There are some known unsafe applications.

UNSAFE APPLICATION(S)
Gender classification · Age classification

UNSAFE USE CASE(S)
This dataset **should not** be used to create gender or age classifiers. The intention of percieved gender and age labels is to capture gender and age presentation as assessed by a third party based on visual cues alone, rather than an individual's self-identified gender or actual age.

CONJUNCTIONAL USE
Safe to use with other datasets

KNOWN CONJUNCTIONAL DATASET(S)
- The data in this dataset can be combined with Open Images V6

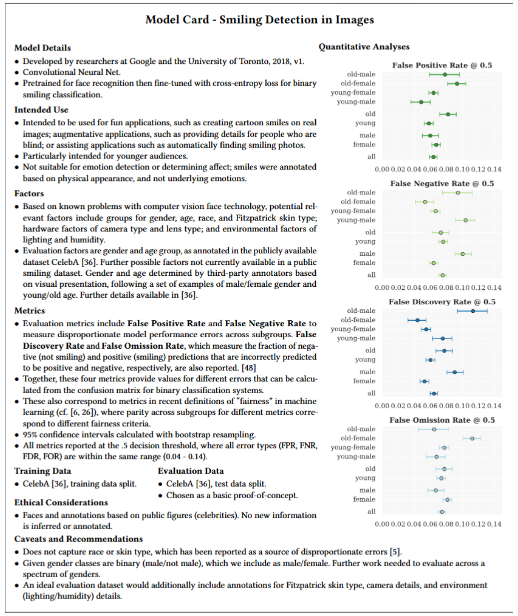
KNOWN CONJUNCTIONAL USES
Analyzing bounding box annotations not annotated under the Open Images V6 procedure.

Example Data Card for Computer Vision Dataset (page 1 of 5) [1]

[1] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA 51 Pages. https://doi.org/10.1145/3531146.3533231
[2] https://sites.research.google/datacardsplaybook/

Model Cards

- Popularized by Mitchell et al. in 2019 [1]
- Used to "standardize ethical practice and reporting"
- For developers to impacted individuals



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

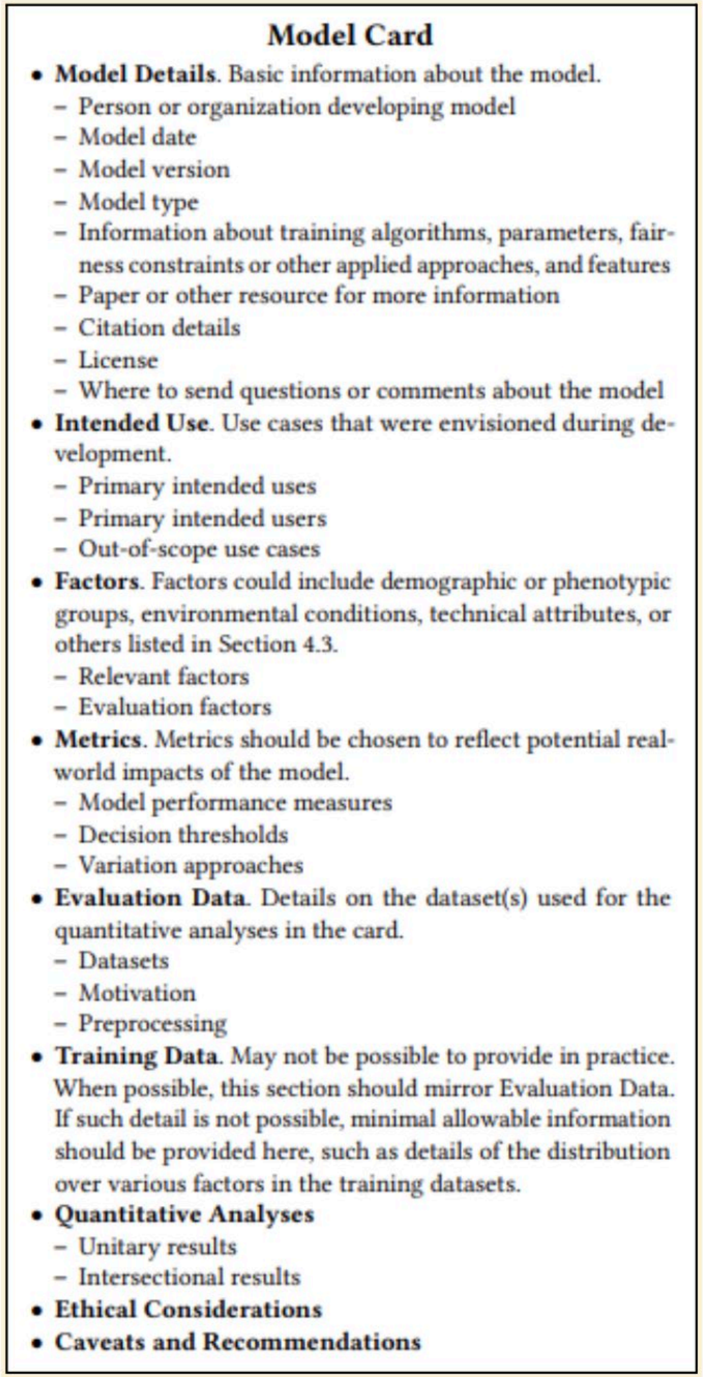
Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

[1] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference

Model Cards in Reality

- Huggingface released a UI to generate Model Cards [1]
- Focused on “discoverability, reproducibility, and sharing”
- Written for other developers

Model Card for {{ model_id | default("Model true) }}

- Model Details
  - Model Description
  - Model Sources [optional]
- Uses
  - Direct Use
  - Downstream Use [optional]
  - Out-of-Scope Use
- **Bias, Risks, and Limitations**
  - Recommendations
- How to Get Started with the Model
- Training Details
  - Training Data
  - Training Procedure
    - Preprocessing [optional]
    - Training Hyperparameters
    - Speeds, Sizes, Times [optional]

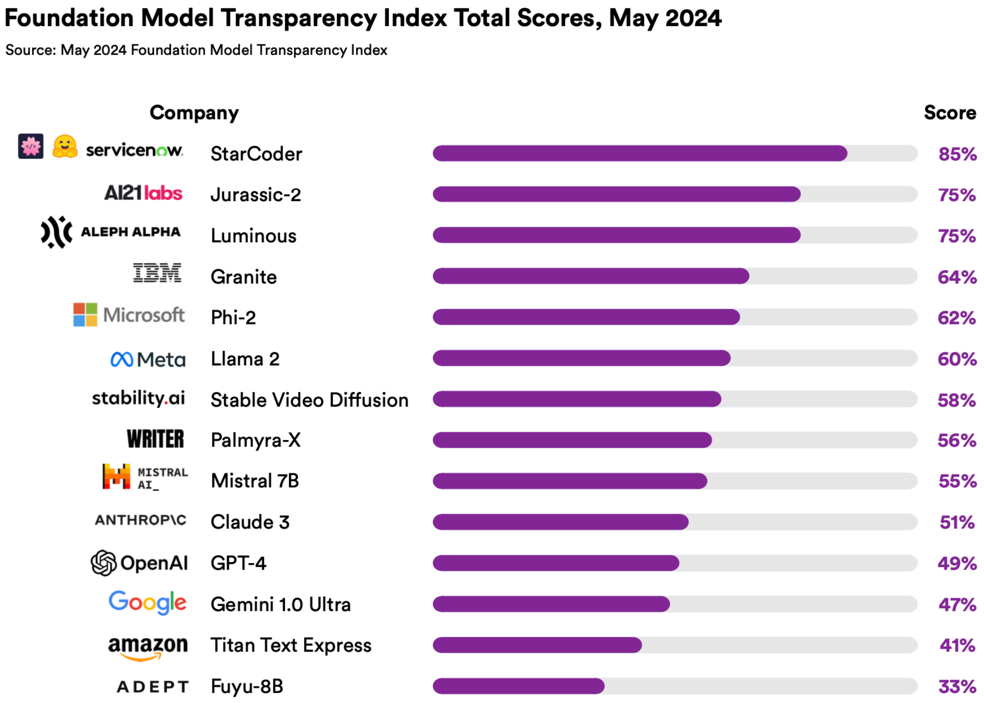
- Evaluation
  - Testing Data, Factors & Metrics
    - Testing Data
    - Factors
    - Metrics
  - Results
    - Summary

Outline of HF Model Card Template [2]

[1] https://huggingface.co/docs/hub/en/model-cards
[2] https://github.com/huggingface/huggingface_hub/blob/main/src/huggingface_hub/templates/modelcard_template.md

Other Transparency Efforts

- Tripod Checklist [1]
- The Foundational Model Transparency Index [2]
- Clear Documentation Framework [3]

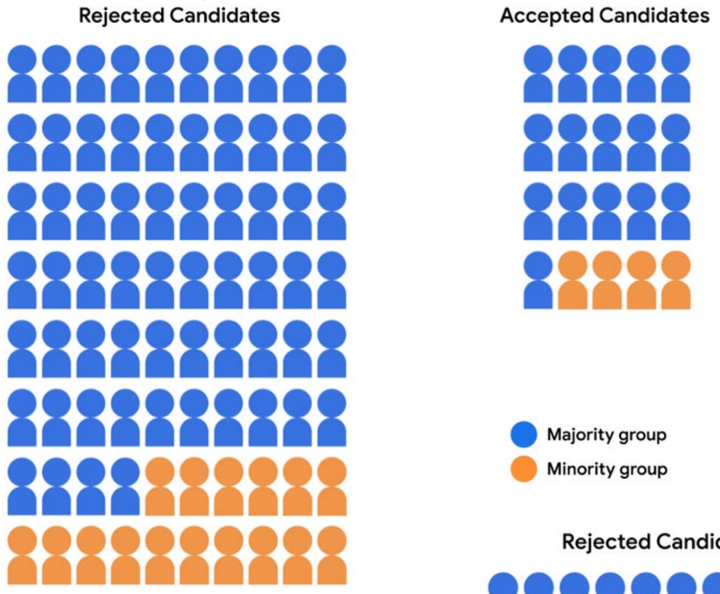


[1] Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.
[2] https://crfm.stanford.edu/fmti/May-2024/index.html
[3] https://shorensteincenter.org/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/

Fairness

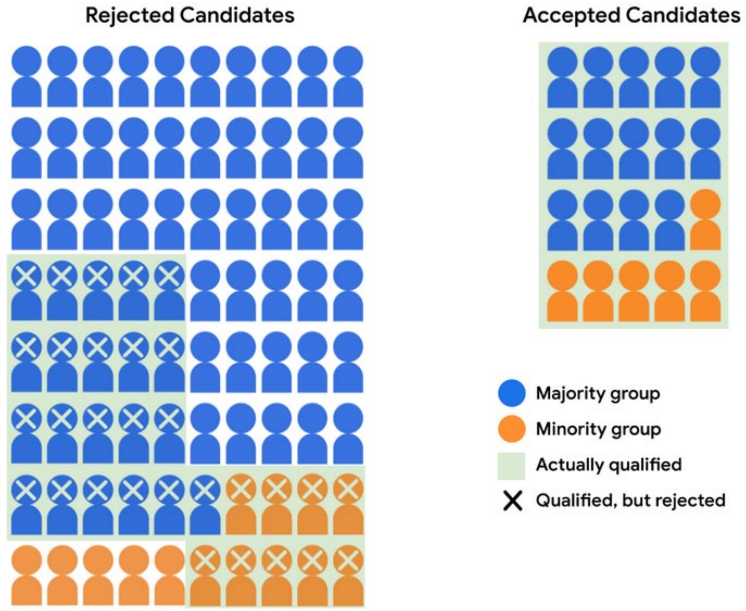
- There's not one way to define and measure fairness
  - Demographic Parity
  - Equal Opportunity
  - Equalized Odds
  - …many more

Demographic Parity



[1]

Equal Opportunity



[1] https://developers.google.com/machine-learning/crash-course/fairness/

Existing Fairness Tools

- AI Fairness 360: https://github.com/Trusted-AI/AIF360
- LinkedIn Fairness Toolkit: https://github.com/linkedin/LiFT
- Fairlearn: https://fairlearn.org
- Causal Fairness Analysis: https://github.com/dplecko/CFA

Our Approach

[1]

Filling in the Gaps

- Transparency written for end users
- Illuminates hidden biases
- No bespoke metrics
- Simplicity!

Model Facts

- Components:
  - How to use it
  - Best by date
  - Nutrition
  - Warnings
- Goal: Provide enough information for users to arrive at their own decisions

Use case

Model/Data relevancy

Metrics for cross model comparison

Metrics for bias evaluation

Model Facts

Application: Brief text string describing use case

Model Type	classification		
Model Train Date	When the model was trained		
Test Data Date	When the test data is from		
Dataset Size	total number of samples		
%Train/%Test	% breakdown		
Accuracy			
	Name	**Raw Score**	**% Over Baseline**
Standard Score	Standard score for the given model type. For cross-model comparisons		
Training Score	Score the model optimized during development		
Demographics	**% in Test Data**	**Standard Score**	**% Target/ Mean,std**
Gender			
female	This section breaks down statistics on the demographics of the data used to test the model. At a minimum, developers should aim to provide information on these categories. This table is extendable to additional demographic groups. Accuracy is reported per the training score name, normalized by each group's distribution. Depending on the model type, either the percent primary target of interest or the mean and standard deviation by demographic group should also be reported.		
male			
trans female			
trans male			
nonbinary			
Age			
<18			
18-24			
25-34			
35-49			
50-64			
64+			
Race			
Asian			
Hispanic			
Black			
White			
Other			

Warning: Any known out of scope use cases, high risk biases, or blind spots (eg, from untested scenarios or missing data).

Where the data is from and who trained the model

How to use Model Facts: The first section, "Application" through "Test Data Date" is to check that this model is relevant and timely for your goals. Use the accuracy "Standard Score" to compare it to other models. Use the demographic breakdown to check for biases in protected attributes (eg, if one race is underrepresented in the "% Test Data" or "% Target" or has a large difference in accuracy compared to the overall model's "Standard Score").

Model Facts generated using https://github.com/jhzsquared/model_facts package

Demo

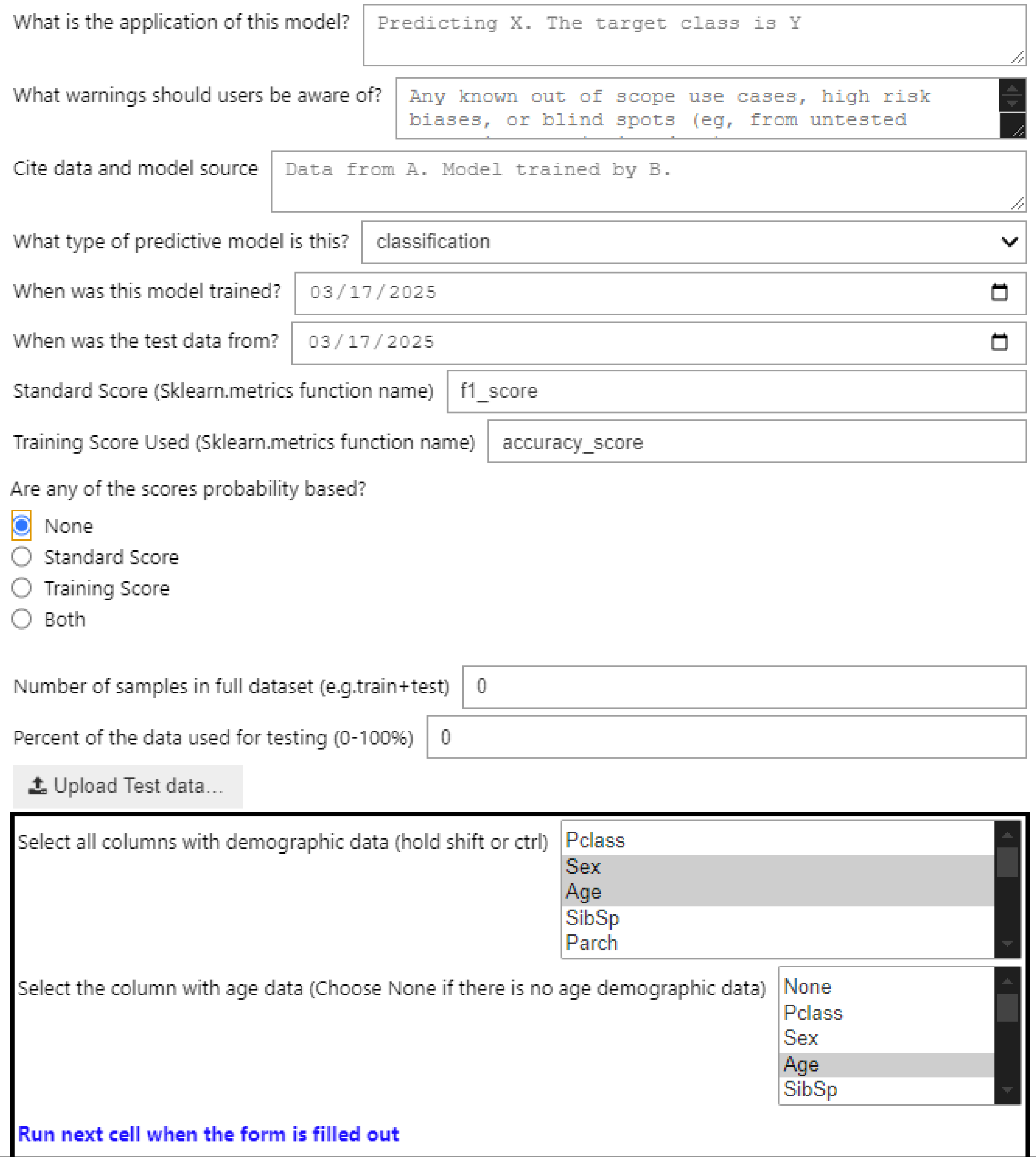
- See example: https://github.com/jhzsquared/model_facts/blob/main/demos/model_facts_titanic.ipynb

How Can I Use It?

- Install modelfacts from pypi

or

- github.com/jhzsquared/modelfacts



What is the application of this model? Predicting X. The target class is Y

What warnings should users be aware of? Any known out of scope use cases, high risk biases, or blind spots (eg, from untested

Cite data and model source Data from A. Model trained by B.

What type of predictive model is this? classification

When was this model trained? 03/17/2025

When was the test data from? 03/17/2025

Standard Score (Sklearn.metrics function name) f1_score

Training Score Used (Sklearn.metrics function name) accuracy_score

Are any of the scores probability based?

- None
- Standard Score
- Training Score
- Both

Number of samples in full dataset (e.g.train+test) 0

Percent of the data used for testing (0-100%) 0

Upload Test data...

Select all columns with demographic data (hold shift or ctrl) Pclass Sex Age SibSp Parch

Select the column with age data (Choose None if there is no age demographic data) None Pclass Sex Age SibSp

Run next cell when the form is filled out

Common Pitfalls

- Bias through awareness vs bias through unawareness
- Not having the right data
- Mis-extrapolation

Model Facts for Gun Violence Research: VOID

Model Facts created from a publication on the Violent Offender Identification Directive Tool used by the Albany Police Department [1]

[1] Wheeler AP, Worden RE, Silver JR. The accuracy of the violent offender identification directive tool to predict future gun violence. *Crim Justice Behav*. 2019;46(5):770-788.

Model Facts

Application: Identify people at very high risk of near-term involvement in gun violence (suspected shooter)

Model Type	classification
Model Train Date	1 January 2012
Test Data Date	1 January 2013
Dataset Size	237232
%Train/%Test	NA/100

Accuracy

	Name	Raw Score	% Over Baseline
Standard Score	f1_score	F1 is not reported	
Training Score	auc	0.939	0.100

Demographics	% in Test Data	Standard Score	% Target
Sex	No demographic information was available or apparent per the published paper		
female			
male			
Age			
<18			

Warnings: The probability of a high-risk individual being involved in gun violence is only around 3% when limiting to the top 1000 scores. Using prior criminal history for estimating risk may propagate any systemic policing biases.

Data and model from a New York Police Department

How to use Model Facts: The first section, "Application" through "Test Data Date" is to check that this model is relevant and timely for your goals. Use the accuracy "Standard Score" to compare it to other models. Use the demographic breakdown to check for biases in protected attributes (eg, if one race is underrepresented in the "% Test Data" or "% Target" or has a large difference in accuracy compared to the overall model's "Standard Score").

Model Facts for Gun Violence Research: COMPAS

- Prompted the "Machine Bias" article by ProPublica
- We re-crunched the ProPublica and NorthPointe numbers

Model Facts

Application: Predicting risk of violent recidivism using COMPAS. The target class is predicting violent recidivism

Using AUC

Model Type	
Model Train Date	
Test Data Date	23 May 2016
Dataset Size	18178
%Train/%Test	NA/100

Accuracy

	Name	Raw Score	% Over Baseline
Standard Score	roc_auc_score	0.648	6.96
Training Score	f1_score	0.172	95.4

Demographics	% in Test Data	Standard Score	% Target
race			
African-American	53.4	0.563	8.64
Asian	0.391	0.548	9.86
Caucasian	33.2	0.520	5.66
Hispanic	7.93	0.519	5.89
Native American	0.314	0.637	10.5
Other	4.70	0.633	6.67
sex			
Female	18.5	0.513	4.55
Male	81.5	0.561	7.99
age			
18-24	23.4	0.585	9.54
25-34	38.7	0.539	8.18
35-49	24.5	0.515	5.85
50-64	12.1	0.514	4.09
64+	1.38	0.500	2.00

Warnings: This model has been demonstrated to propagate biases by ProPublica. Its creators claim this model is unbiased, under the predictive parity paradigm using AUC. Without a clear definition of fairness, it should not be used in decision making

Data from Broward County, Florida https://github.com/propublica/compas-analysis/tree/master. Model created by Northpointe

How to use Model Facts: The first section, "Application" through "Test Data Date" is to check that this model is relevant and timely for your goals. Use the accuracy "Standard Score" to compare it to other models. Use the demographic breakdown to check for biases in protected attributes (eg, if one race is underrepresented in the "% Test Data" or "% Target" or has a large difference in accuracy compared to the overall model's "Standard Score").

[1] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[2] https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Model Facts for Gun Violence Research: COMPAS

Honest transparency:

- In sharing weaknesses
- Or in misapplication of context and metrics

[1] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[2] https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Model Facts

Application: Predicting risk of violent recidivism using COMPAS. The target class is predicting violent recidivism

Using F-1

Model Type			
Model Train Date			
Test Data Date	23 May 2016		
Dataset Size	18178		
%Train/%Test	NA/100		
Accuracy			
	Name	**Raw Score**	**% Over Baseline**
Standard Score	f1_score	0.172	95.4
Training Score	roc_auc_score	0.648	6.96
Demographics	**% in Test Data**	**Standard Score**	**% Target**
race			
African-American	53.4	0.191	8.64
Asian	0.391	0.182	9.86
Caucasian	33.2	0.0946	5.66
Hispanic	7.93	0.0985	5.89
Native American	0.314	0.364	10.5
Other	4.70	0.340	6.67
sex			
Female	18.5	0.0732	4.55
Male	81.5	0.182	7.99
age			
18-24	23.4	0.220	9.54
25-34	38.7	0.156	8.18
35-49	24.5	0.0832	5.85
50-64	12.1	0.0615	4.09
64+	1.38	0.00	2.00

Warnings: This model has been demonstrated to propagate biases by ProPublica. Its creators claim this model is unbiased, under the predictive parity paradigm using AUC. Without a clear definition of fairness, it should not be used in decision making

Data from Broward County, Florida https://github.com/propublica/compas-analysis/tree/master. Model created by Northpointe

How to use Model Facts: The first section, "Application" through "Test Data Date" is to check that this model is relevant and timely for your goals. Use the accuracy "Standard Score" to compare it to other models. Use the demographic breakdown to check for biases in protected attributes (eg, if one race is underrepresented in the "% Test Data" or "% Target" or has a large difference in accuracy compared to the overall model's "Standard Score").

Model Facts

Application: Predicting risk of violent recidivism using COMPAS. The target class is predicting violent recidivism

Using AUC

Model Type			
Model Train Date			
Test Data Date	23 May 2016		
Dataset Size	18178		
%Train/%Test	NA/100		
Accuracy			
	Name	**Raw Score**	**% Over Baseline**
Standard Score	roc_auc_score	0.648	6.96
Training Score	f1_score	0.172	95.4
Demographics	**% in Test Data**	**Standard Score**	**% Target**
race			
African-American	53.4	0.563	8.64
Asian	0.391	0.548	9.86
Caucasian	33.2	0.520	5.66
Hispanic	7.93	0.519	5.89
Native American	0.314	0.637	10.5
Other	4.70	0.633	6.67
sex			
Female	18.5	0.513	4.55
Male	81.5	0.561	7.99
age			
18-24	23.4	0.585	9.54
25-34	38.7	0.539	8.18
35-49	24.5	0.515	5.85
50-64	12.1	0.514	4.09
64+	1.38	0.500	2.00

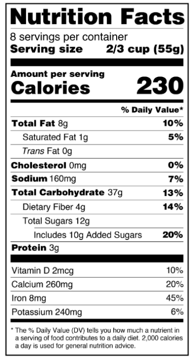
Warnings: This model has been demonstrated to propagate biases by ProPublica. Its creators claim this model is unbiased, under the predictive parity paradigm using AUC. Without a clear definition of fairness, it should not be used in decision making

Data from Broward County, Florida https://github.com/propublica/compas-analysis/tree/master. Model created by Northpointe

How to use Model Facts: The first section, "Application" through "Test Data Date" is to check that this model is relevant and timely for your goals. Use the accuracy "Standard Score" to compare it to other models. Use the demographic breakdown to check for biases in protected attributes (eg, if one race is underrepresented in the "% Test Data" or "% Target" or has a large difference in accuracy compared to the overall model's "Standard Score").

Remaining Gaps

- Only for predictive modeling (currently)
- User experience
- Education
- How do we set the standard?
- Transparency vs responsibility



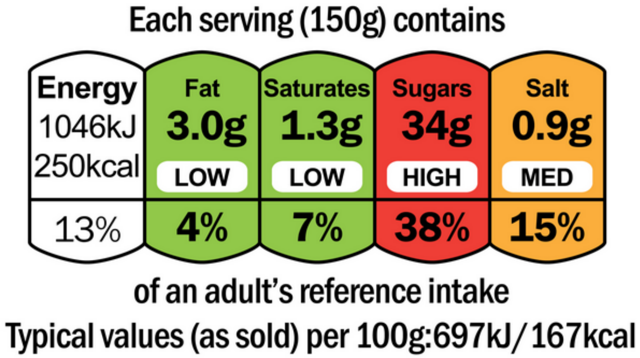
Nutrition Facts

8 servings per container
Serving size 2/3 cup (55g)

Amount per serving
Calories 230

	% Daily Value*
Total Fat 8g	10%
Saturated Fat 1g	5%
Trans Fat 0g	
Cholesterol 0mg	0%
Sodium 160mg	7%
Total Carbohydrate 37g	13%
Dietary Fiber 4g	14%
Total Sugars 12g	
Includes 10g Added Sugars	20%
Protein 3g	
Vitamin D 2mcg	10%
Calcium 260mg	20%
Iron 8mg	45%
Potassium 240mg	6%

* The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.



[1] https://www.nutrition.org.uk/nutritional-information/sugar/

Concluding Call For:

- Transparency for your customers
- Clear communication
- Standardization of transparency practices
- Coordinating AI/data literacy education with industry standards

Thanks! Questions?

- Paper: Jessica Zhu, Michel Cukier, Joseph Richardson, Nutrition facts, drug facts, and model facts: putting AI ethics into practice in gun violence research, *Journal of the American Medical Informatics Association*, Volume 31, Issue 10, October 2024, Pages 2414–2421, https://doi.org/10.1093/jamia/ocae102
- Python package: https://pypi.org/project/modelfacts/
- Source code and examples: https://github.com/jhzsquared/model_facts
- LinkedIn: https://www.linkedin.com/in/jessicazhu28/